CS 112: Computer System Modeling Fundamentals

Prof. Jenn Wortman Vaughan

May 5, 2011

Lecture 11

Reminders & Announcements

- Homework 2 grades have been posted in Gradebook
- Homework 3 is due this Tuesday

Before the Midterm

- Bounds on probabilities
  - Markov Inequality
  - Chebyshev Inequality

- Applying these bounds to sample means
  - The (weak) law of large numbers

Law of Large Numbers

Theorem: Let X_1, X_2, … be independent identically distributed random variables with mean μ. For every $\varepsilon > 0$,

$$P\left(\left| \frac{X_1 + \ldots + X_n}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as} \quad n \to \infty$$

Translation: as the size of our sample gets very large, the probability that the sample mean is very close to the true mean goes to 1.

Today…

- Normal or Gaussian random variables
- The Central Limit Theorem

- New topic: Statistical inference

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

just for normalization

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Things to notice:

- Maximized when $x = 0$

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Things to notice:

- Maximized when $x = 0$
- Symmetric around 0

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Things to notice:

- Maximized when $x = 0$
- Symmetric around 0
- Drops exponentially fast

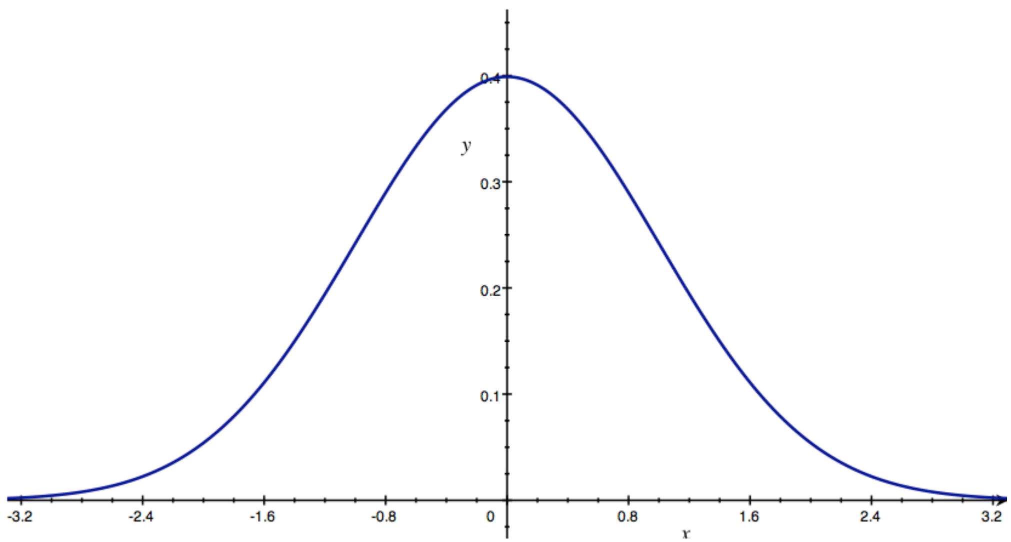
Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Things to notice:

- Maximized when $x = 0$
- Symmetric around 0
- Drops exponentially fast



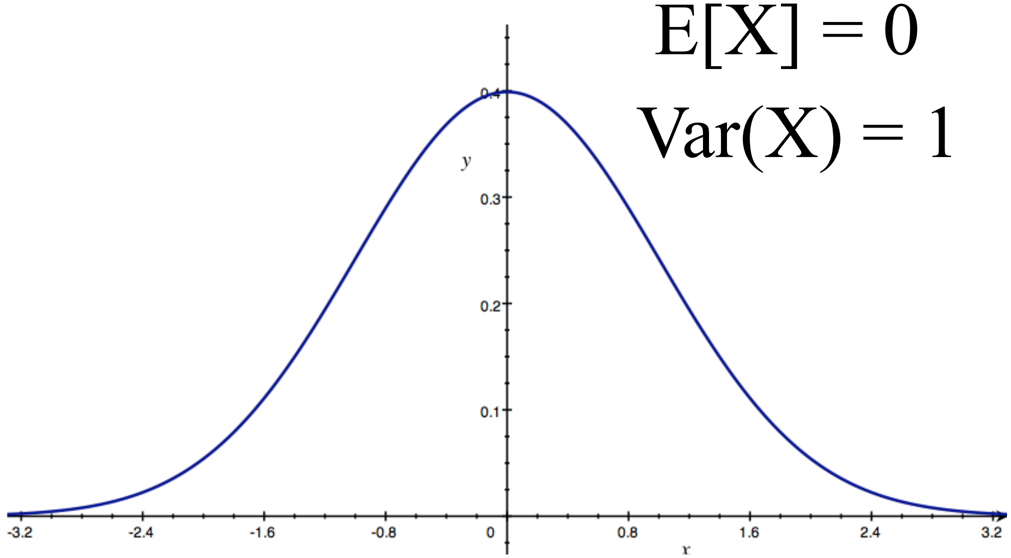
Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

E[X] = 0

Var(X) = 1



Things to notice:

- Maximized when $x = 0$
- Symmetric around 0
- Drops exponentially fast

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

first term still just normalization

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Things to notice:

- Maximized when $x = \mu$

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Things to notice:

- Maximized when $x = \mu$
- Symmetric around μ

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Things to notice:

- Maximized when $x = \mu$
- Symmetric around μ
- Drops exponentially fast

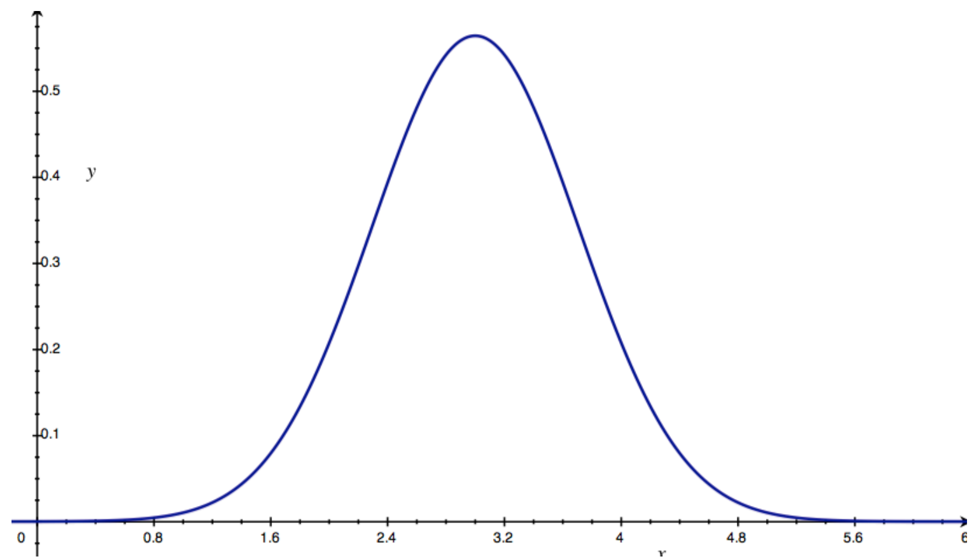
Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Things to notice:

- Maximized when $x = \mu$
- Symmetric around μ
- Drops exponentially fast



Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

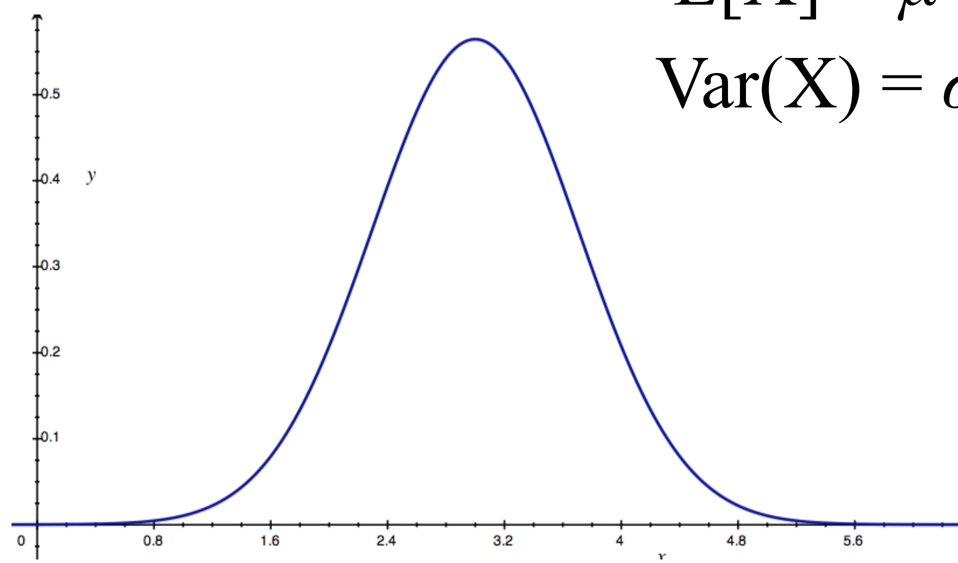
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$\mathrm{E}[\mathrm{X}] = \mu$

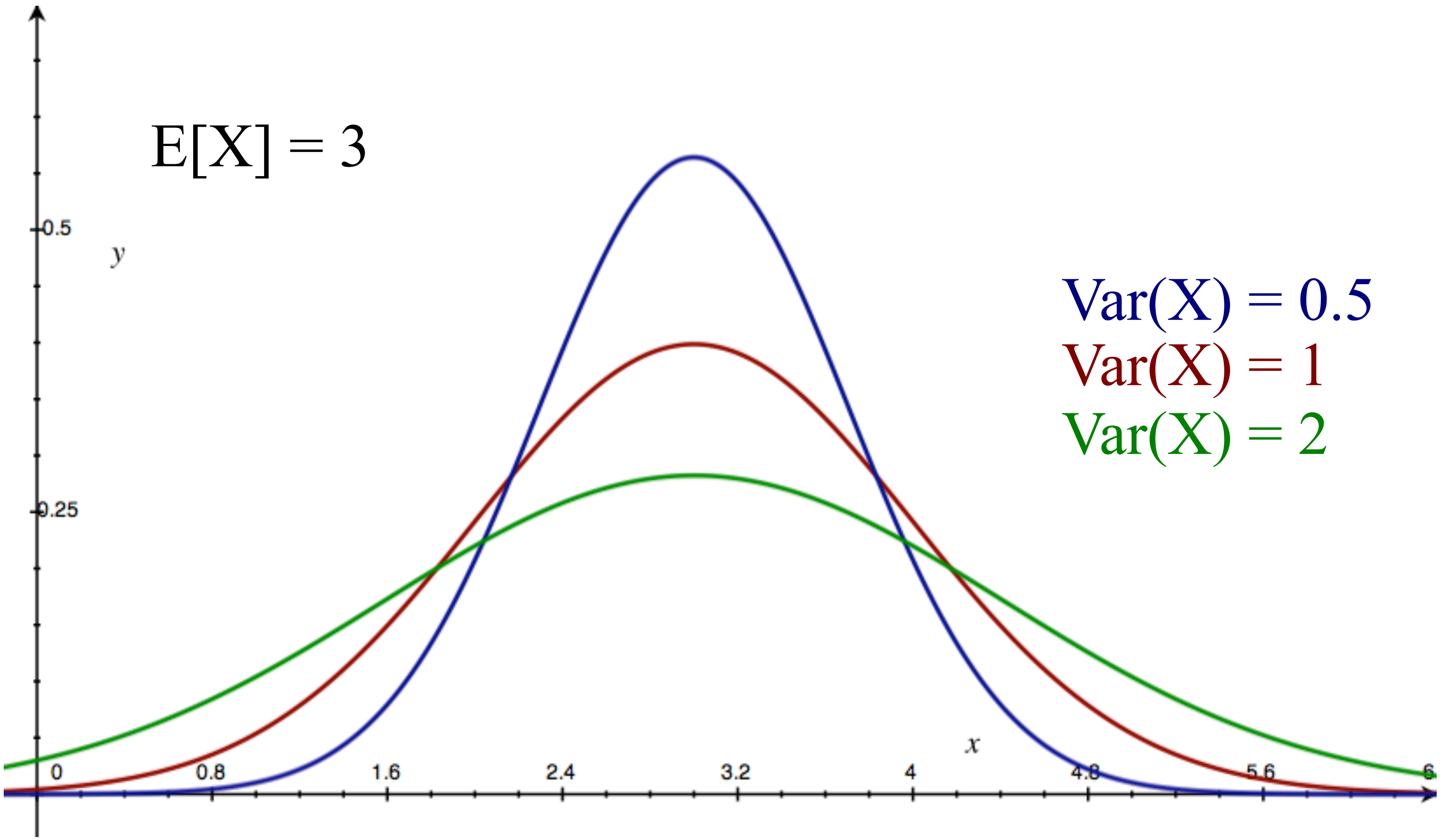
$\mathrm{Var}(\mathrm{X}) = \sigma^2$

Things to notice:

- Maximized when $x = \mu$
- Symmetric around μ
- Drops exponentially fast



Normal Random Variables



Normal Random Variables

Normality is preserved by linear transformations:

- Let X be a normal random variable with mean μ and standard deviation σ
- Let Y = aX + b, for any $a \neq 0$ and any b

Normal Random Variables

Normality is preserved by linear transformations:

- Let X be a normal random variable with mean μ and standard deviation σ
- Let $Y = aX + b$, for any $a \neq 0$ and any b

- Then Y is also normal with

$$E[Y] = a\mu + b$$

$$\mathrm{var}(Y) = (a\sigma)^2$$

Back to the sample mean..

- Independent and identically distributed $X_1, X_2, X_3, \ldots$
- $S_n = X_1 + X_2 + \ldots + X_n$
- $M_n = (1/n)\, S_n$

We know that M_n converges as n gets large

What about S_n?

Back to the sample mean..

- Independent and identically distributed $X_1, X_2, X_3, \ldots$
- $S_n = X_1 + X_2 + \ldots + X_n$
- $M_n = (1/n)\ S_n$

We know that M_n converges as n gets large

What about S_n?

We'll come back to this question, but first, we'll examine a slightly different quantity…

The Central Limit Theorem

Theorem: Let X_1, X_2, … be independent identically distributed random variables with mean μ and standard deviation σ, and define

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

Then for every value z,

$$\lim_{n \to \infty} P(Z_n \leq z) = \Phi(z)$$

where Φ is the CDF of the standard normal RV

Implications

- Suppose Z_n is a normal random variable. What does that tell us about the sum $S_n = X_1 + X_2 + \ldots + X_n$?

Implications

- Suppose Z_n is a normal random variable. What does that tell us about the sum $S_n = X_1 + X_2 + \ldots + X_n$?

- What if Z_n is *approximately* normal?

Normal Approximation

Let $S_n = X_1 + X_2 + \ldots + X_n$, where the X_i are independent identically distributed random variables with mean μ and standard deviation σ. If n is large, then for any c,

$$P(S_n \leq c) = P\left(Z_n \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

where Φ is the CDF of the standard normal RV

Normal Approximation

Let $S_n = X_1 + X_2 + \ldots + X_n$, where the X_i are independent identically distributed random variables with mean μ and standard deviation σ. If n is large, then for any c,

$$P(S_n \leq c) = P\left(Z_n \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

where Φ is the CDF of the standard normal RV

(Values of Φ can be looked up in tables, e.g., in the textbook)

Normal Approximation

We run a program 100 times. The runtimes are independent random variables uniformly distributed between 5 and 50 seconds. What is the probability that the *total* runtime will exceed 3000 seconds?

Polling Again

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

What is the probability that our estimate differs from the true approval rating by more than ε?

Polling Again

Suppose we would like to estimate the president's approval rating. We ask n random voters whether or not they approve of the president, and use the fraction of voters who say that they approve as our estimate.

What is the probability that our estimate differs from the true approval rating by more than ε?

Can use the fact that M_n is normal to get a better bound than we were able to get with Chebyshev

(Try this on your own, or read example 5.11 in the book.)

New Topic: Statistical Inference

So far, we have (almost always) assumed the existence of a probabilistic model obeying the laws of probability.

The questions we've asked have a unique right answer.

New Topic: Statistical Inference

So far, we have (almost always) assumed the existence of a probabilistic model obeying the laws of probability.

The questions we've asked have a unique right answer.

- A fair die is rolled three times. What is the probability that all rolls are more than three?

New Topic: Statistical Inference

So far, we have (almost always) assumed the existence of a probabilistic model obeying the laws of probability.

The questions we've asked have a unique right answer.

- A fair die is rolled three times. What is the probability that all rolls are more than three?
- If the time before a hard disk fails is modeled as an exponential random variable with mean λ, how likely is it to fail in the first two years?

New Topic: Statistical Inference

In statistical inference, we are given only observations.

There may not always be a single “right” answer…

New Topic: Statistical Inference

In statistical inference, we are given only observations.

There may not always be a single "right" answer…

- Based on a collection of old email, how likely is it that this new email is spam?

New Topic: Statistical Inference

In statistical inference, we are given only observations.

There may not always be a single "right" answer…

- Based on a collection of old email, how likely is it that this new email is spam?
- Based on polling data, what fraction of the population approves of this candidate? (We discussed one way to estimate this value…)

New Topic: Statistical Inference

In statistical inference, we are given only observations.

There may not always be a single "right" answer…

- Based on a collection of old email, how likely is it that this new email is spam?
- Based on polling data, what fraction of the population approves of this candidate? (We discussed one way to estimate this value…)

For the next few classes, we will discuss different methods and techniques that can be used to answer these questions.

Types of Inference

Hypothesis testing: Decide which of two or more hypotheses is more likely to true based on some data.

- Determine whether an email containing a particular set of words is more likely to be spam or not spam
- Given a student's test score, decide if he studied or not

Types of Inference

Hypothesis testing: Decide which of two or more hypotheses is more likely to true based on some data.

- Determine whether an email containing a particular set of words is more likely to be spam or not spam
- Given a student's test score, decide if he studied or not

Parameter estimation: Model is fully specified except some unknown parameters we need to estimate.

- Estimate the bias of a coin from a sequence of flips
- Estimate the fraction of the population who prefers candidate A to candidate B based on polling data

Hypothesis Testing

Let D be the event that we observed some particular data

- D = event that I observed an email containing the words "ca$h" and "viagra"

Hypothesis Testing

Let D be the event that we observed some particular data

- D = event that I observed an email containing the words "ca$h" and "viagra"

Let $H_1, \ldots, H_k$ be disjoint and exhaustive events representing hypotheses we are choosing among

- H_1 = event that the email is spam
- H_2 = event that the email is not spam

Hypothesis Testing

Let D be the event that we observed some particular data

- D = event that I observed an email containing the words "ca$h" and "viagra"

Let $H_1, \ldots, H_k$ be disjoint and exhaustive events representing hypotheses we are choosing among

- H_1 = event that the email is spam
- H_2 = event that the email is not spam

What is the most likely hypothesis given the data?

Maximum Likelihood

- Suppose that we know (or can compute) the probability $P(D \mid H_i)$ of observing data D for each hypothesis H_i

Maximum Likelihood

- Suppose that we know (or can compute) the probability $P(D \mid H_i)$ of observing data D for each hypothesis H_i

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i \; P(D \mid H_i)$$

Maximum Likelihood

There are two boxes of cookies. One contains half chocolate chip cookies and half oatmeal raison cookies. The other contains one third chocolate chip cookies and two thirds oatmeal raison. I select a box and pull a random cookie from it. You observe that the cookie is chocolate chip.

Which box is most likely to be the one I chose from?

Maximum Likelihood

There are two boxes of cookies. One contains half chocolate chip cookies and half oatmeal raison cookies. The other contains one third chocolate chip cookies and two thirds oatmeal raison. I select a box and pull a random cookie from it. You observe that the cookie is chocolate chip.

Which box is most likely to be the one I chose from?

- D = event that I chose a chocolate chip cookie
- $P(D|H_1) = 0.5$, $P(D|H_2) = 0.33$
- $H^{ML} = H_1$

Maximum Likelihood

There are two boxes of cookies. One contains half chocolate chip cookies and half oatmeal raison cookies. The other contains one third chocolate chip cookies and two thirds oatmeal raison. I select a box and pull a random cookie from it. You observe that the cookie is chocolate chip.

Which box is most likely to be the one I chose from?

What if you know that box 2 is lying out on the table, while box 1 is put away?

Bayesian Reasoning

- Suppose that we know (or can compute) the probability $P(D|H_i)$ of observing data D for each hypothesis H_i

Bayesian Reasoning

- Suppose that we know (or can compute) the probability $P(D|H_i)$ of observing data D for each hypothesis H_i

- Suppose that we have a prior belief $P(H_i)$ about how likely each hypothesis H_i is (this can be a subjective probability)

Bayesian Reasoning

- Suppose that we know (or can compute) the probability $P(D|H_i)$ of observing data D for each hypothesis H_i
- Suppose that we have a prior belief $P(H_i)$ about how likely each hypothesis H_i is (this can be a subjective probability)
- We can use Bayes' rule to compute a posterior belief about how likely each hypothesis is given the data we observed:

Bayesian Reasoning

- Suppose that we know (or can compute) the probability $P(D|H_i)$ of observing data D for each hypothesis H_i

- Suppose that we have a prior belief $P(H_i)$ about how likely each hypothesis H_i is (this can be a subjective probability)

- We can use Bayes' rule to compute a posterior belief about how likely each hypothesis is given the data we observed:

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)} = \frac{P(D \mid H_i)P(H_i)}{\sum_j P(D \mid H_j)P(H_j)}$$

Maximum a Posteriori

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i \; P(D \mid H_i)$$

Maximum a Posteriori

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i\ P(D \mid H_i)$$

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i\ P(H_i \mid D)$$

Maximum a Posteriori

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i \; P(D \mid H_i)$$

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i \; P(H_i \mid D) = \text{argmax}_i \; P(D \mid H_i) \, P(H_i)$$

Maximum a Posteriori

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i\ P(D \mid H_i)$$

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i\ P(H_i \mid D) = \text{argmax}_i\ P(D \mid H_i)\ P(H_i)$$

When are these the same?

(we ran out of time and stopped here in class, but we'll continue with this overview of the inference part of the course in the next lecture)

Bayesian Reasoning

There are two boxes of cookies. One contains half chocolate chip cookies and half oatmeal raison cookies. The other contains one third chocolate chip cookies and two thirds oatmeal raison. I select a box and pull a random cookie from it. You observe that the cookie is chocolate chip.

If you know that box 2 is on the table and box 1 is put away, which box is most likely to be the one I chose from?

Bayesian Reasoning

There are two boxes of cookies. One contains half chocolate chip cookies and half oatmeal raison cookies. The other contains one third chocolate chip cookies and two thirds oatmeal raison. I select a box and pull a random cookie from it. You observe that the cookie is chocolate chip.

If you know that box 2 is on the table and box 1 is put away, which box is most likely to be the one I chose from?

- $P(H_1) = 0.1$, $P(H_2) = 0.9$ (for example..)
- $H^{MAP} = H_2$

Next Couple Weeks…

- Classical statistical inference
  - Justification for maximum likelihood
  - Naive Bayes classifier using maximum likelihood estimates (which you will implement for homework 4!)
  - Confidence bounds

- Bayesian statistical inference
  - Priors and posteriors
  - MAP estimation